Stacking models for nearly optimal link prediction in complex networks

Amir Ghasemian[a,b,c,1], Homa Hosseinmardi[b], Aram Galstyan[b], Edoardo M. Airoldi[c,d], and Aaron Clauset[a,e,f,1]

[a]Department of Computer Science, University of Colorado, Boulder, CO 80309; [b]Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292; [c]Department of Statistics, Harvard University, Cambridge, MA 02138; [d]Department of Statistical Science, Fox School of Business, Temple University, Philadelphia, PA 19122; [e]BioFrontiers Institute, University of Colorado, Boulder, CO 80303; and [f]Santa Fe Institute, Santa Fe, NM 87501

Edited by Luís A. Nunes Amaral, Northwestern University, Evanston, IL, and accepted by Editorial Board Member Simon A. Levin August 6, 2020 (received for review September 2, 2019)

Most real-world networks are incompletely observed. Algorithms that can accurately predict which links are missing can dramatically speed up network data collection and improve network model validation. Many algorithms now exist for predicting missing links, given a partially observed network, but it has remained unknown whether a single best predictor exists, how link predictability varies across methods and networks from different domains, and how close to optimality current methods are. We answer these questions by systematically evaluating 203 individual link predictor algorithms, representing three popular families of methods, applied to a large corpus of 550 structurally diverse networks from six scientific domains. We first show that individual algorithms exhibit a broad diversity of prediction errors, such that no one predictor or family is best, or worst, across all realistic inputs. We then exploit this diversity using network-based metalearning to construct a series of "stacked" models that combine predictors into a single algorithm. Applied to a broad range of synthetic networks, for which we may analytically calculate optimal performance, these stacked models achieve optimal or nearly optimal levels of accuracy. Applied to real-world networks, stacked models are superior, but their accuracy varies strongly by domain, suggesting that link prediction may be fundamentally easier in social networks than in biological or technological networks. These results indicate that the state of the art for link prediction comes from combining individual algorithms, which can achieve nearly optimal predictions. We close with a brief discussion of limitations and opportunities for further improvements.

link prediction | stacking | networks | metalearning | near optimality

Networks provide a powerful abstraction for representing the structure of complex social, biological, and technological systems. However, data on most real-world networks are incomplete. For instance, social connections among people may be sampled, intentionally hidden, or simply unobservable (1, 2); interactions among genes, cells, or species must be observed or inferred by expensive experiments (3, 4); and connections mediated by a particular technology omit all off-platform interactions (2, 5). The presence of such "missing links" can, depending on the research question, dramatically alter scientific conclusions when analyzing a network's structure or modeling its dynamics.

Methods that accurately predict which observed pairs of unconnected nodes should, in fact, be connected have broad utility. For instance, they can improve the accuracy of predictions of future network structure and minimize the use of scarce experimental or network measurement resources (6, 7). Moreover, the task of link prediction itself has become a standard for evaluating and comparing models of network structure (8, 9), playing a role in networks that is similar to that of cross-validation in traditional statistical learning (10, 11). Hence, by helping to select more accurate network models (8), methods for link prediction can shed light on the organizing principles of complex systems of all kinds.

However, predicting missing links is a statistically hard problem. Most real-world networks are relatively sparse, and the number of unconnected pairs in an observed network—each a potential missing link—grows quadratically, like $O(n^2)$ for a network with n nodes when the number of connected pairs or edges m grows linearly, like $O(n)$. The probability of correctly choosing by chance a missing link is thus only $O(1/n)$—an impractically small chance even for moderate-sized systems (12). Despite this baseline difficulty, a plethora of link prediction methods exists (3, 13, 14), embodied by the three main families we study here: 1) topological methods (15, 16), which utilize network measures like node degrees, the number of common neighbors, and the length of a shortest path; 2) model-based methods (8, 12, 17), such as the stochastic block model, its variants, and other models of community structure; and 3) embedding methods (18, 19), which project a network into a latent space and predict links based on the induced proximity of its nodes.

A striking feature of this array of methods is that all appear to work relatively well (8, 15, 18). However, systematic comparisons are lacking, particularly of methods drawn from different families, and most empirical evaluations are based on relatively small numbers of networks. As a result, the general accuracy of different methods remains unclear, and we do not know whether different methods, or families, are capturing the same underlying signatures of "missingness." For instance, is there a single best

Significance

Networks are a powerful tool for modeling complex biological and social systems. However, most networks are incomplete, and missing connections can negatively affect scientific analyses. Today, many algorithms can predict missing connections, but it is unknown how accuracy varies across algorithms and networks and whether link predictability varies across scientific domains. Analyzing 203 link prediction algorithms applied to 550 diverse real-world networks, we show that no predictor is best or worst overall. We then combine these many predictors into a single state-of-the-art algorithm that achieves nearly optimal performance on both synthetic networks with known optimality and real-world networks. Not all networks are equally predictable, however, and we find that social networks are easiest, while biological and technological networks are hardest.

Author contributions: A. Ghasemian, H.H., A. Galstyan, E.M.A., and A.C. designed research; A. Ghasemian and H.H. performed research; A. Ghasemian and H.H. analyzed data; and A. Ghasemian and A.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. L.A.N.A. is a guest editor invited by the Editorial Board.

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: amir.ghasemianlangroodi@colorado.edu or aaron.clauset@colorado.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1914950117/-/DCSupplemental.

First published September 4, 2020.

method or family for all circumstances? If not, then how does missing link predictability vary across methods and scientific domains (e.g., in social vs. biological networks) or across network scales? Additionally, how close to optimality are current methods?

Here, we answer these questions using a large corpus of 550 structurally and scientifically diverse real-world networks and 203 missing link predictors drawn from three large methodological families. To begin, we present broad empirical evidence that individual methods for link prediction exploit different underlying signals of missingness. This finding implicates the practical relevance of the No Free Lunch theorem (20, 21) for link prediction in general, by showing empirically that no known method performs best or worst across realistic inputs.

To exploit this empirical diversity of errors, we then adopt the metalearning approach (22–24) of model stacking (25), which we adapt here to the setting of network (relational) data. We then argue that these stacked models yield nearly optimal predictions of missing links in sparse, realistic networks, based on three lines of evidence: 1) evaluations on synthetic data with known structure and optimal performance, 2) tests using real-world networks across scientific domains and network scales, and 3) tests of sufficiency and saturation using subsets of methods.

Across these tests, we find that model stacking is nearly always the best method on average on held-out links and that nearly optimal performance can be constructed using model-based methods, topological methods, or a mixture of the two. These stacked models achieve their high performance by combining individual predictors into an effective Bayesian model of the general structural diversity of real-world networks, and their performance improves with network size. Finally, we find that missing links are generally easiest to predict in social networks, where most methods perform well, and hardest in biological and technological networks. We conclude by discussing limitations and opportunities for further improvement of these results. To facilitate these future directions of work, we make available code for training a stacked model using topological predictors on an arbitrary network and the large corpus of networks used here.

Methods and Materials

As a general setting, we imagine an unobserved simple network G with a set of E pairwise connections among a set of V nodes, with sizes m and n, respectively. Of these, a subset $E' \subset E$ of connections is observed, chosen by some function f. Our task is to accurately guess, based only on the pattern of observed edges E', which unconnected pairs $X = V \times V - E'$ are in fact among the missing links $Y = E - E'$. A link prediction method defines a score function over these unconnected pairs $i, j \in X$ so that better-scoring pairs are more likely to be missing links (15). In a supervised setting, the particular function that combines input predictors to produce a score is learned from the data. We evaluate the accuracy of such predictions using the standard area under the curve (AUC) statistic, which provides an easily interpretable and context-agnostic measure of a method's ability to distinguish a missing link $i, j \in Y$ (a true positive) from a nonedge $X - Y$ (a true negative) (12) and facilitates comparison with other link prediction methods in the literature. We also report precision and recall metrics at a threshold that maximizes the F measure for each network.

The most common approach to predict missing links constructs a score function from network statistics of each unconnected node pair (15). We study 42 of these topological predictors, which include predictions based on node degrees, common neighbors, random walks, and node and edge centralities, among others (*SI Appendix*, Table S1). Models of large-scale network structure are also commonly used for link prediction. We study 11 of these model-based methods (8), which either estimate a parametric probability $\Pr(i \to j \mid \theta)$ that a node pair is connected (12), given a decomposition of a network into communities, or predict a link as missing if it would improve a measure of community structure (15) (*SI Appendix*, Table S2). Close proximity of an unconnected pair, after embedding a network's nodes into a latent space (19), is a third common approach to link prediction. We study 150 of these embedding-based predictors, derived from two popular graph embedding algorithms and six notions of distance or similarity in the latent space. In total, we consider 203 features of node pairs, some of which are the output of existing link prediction algorithms, while others are numerical features derived from the network structure. For our purposes, each is considered a missing link "predictor." A lengthier description of these 203 methods, and the three methodological families they represent, is given in *SI Appendix*, section A. For brevity, we use abbreviations to identify individual methods in the main text. A complete listing of methods and their abbreviations is given in *SI Appendix*, section A.

Metalearning is a powerful class of ensemble methods with machine learning that can learn from data how to combine individual predictors into a single, more accurate algorithm (23, 26). In particular, stacked generalization (25) combines predictors by learning a supervised model of input characteristics and the corresponding errors made by individual predictors. Model stacking thus enriches the space of hypotheses by treating a set of predictors as a panel of experts and then learning the kinds of questions each is most expert at answering. From a Bayesian perspective, stacking combines models so as to asymptotically minimize posterior loss (27, 28). Bayesian model averaging (29) is another common ensemble method, which operates more as a model selection technique than as a model combination method (30–32), and is particularly useful when model predictions are probabilistic. Model averaging has previously been used for link prediction, by sampling an ensemble of hierarchical random graphs (12) or stochastic block models (17). In contrast, model stacking's ability to flexibly combine arbitrary component predictors to learn stacked weights that asymptotically minimize a posterior expected loss makes it an attractive approach to investigate the broad variety of link prediction algorithms considered here. Moreover, stacked models can be strictly more accurate than their component predictors (25), making them appropriate for hard problems like link prediction (33), but only if those predictors make distinct errors and are sufficiently diverse in the signals they exploit (27). To apply model stacking to link prediction, we first adapt its form to the setting of network (relational) data (*SI Appendix*, section A).

We evaluate individual prediction methods, and their stacked generalizations, using two types of network data. The first is a set of synthetic networks with known structure that varies along three dimensions: 1) the degree distribution's variability, being low (Poisson), medium (Weibull), or high (power law); 2) the number of "communities" or modules $k \in \{1, 2, 4, 16, 32\}$; and 3) the fuzziness of the corresponding community boundaries ϵ, being low, medium, or high. These synthetic networks thus range from homogeneous to heterogeneous random graphs, from no modules to many modules, and from weakly to strongly modular structure (*SI Appendix*, section B and Table S3). Moreover, because the data-generating process for these networks is known, we exactly calculate the optimal accuracy that any link prediction method could achieve, as a reference point (*SI Appendix*, section B).

The second is a large and structurally diverse corpus of 550 real-world networks, which is a slight expansion of the popular CommunityFitNet corpus (8). It includes social (23%), biological (32%), economic (23%), technological (12%), information (3%), and transportation (7%) networks, and these networks span three orders of magnitude in size (*SI Appendix*, section C and Fig. S1). It is by far the largest and most diverse empirical link prediction benchmark, which enables the comparison of methods across scientific domains.

Finally, our evaluations assume a missingness function f that samples edges uniformly at random from E so that each edge $(i, j) \in E$ is observed with probability α. This choice presents a hard test, as f is independent of both observed edges and metadata. Other models of f (e.g., in which missingness correlates with edge or node characteristics) may better capture particular scientific settings and are left for future work. Our results thus provide a general, application-agnostic assessment of link predictability and method performance. In cases of supervised learning, we train a method using fivefold cross-validation by choosing as positive examples a subset of edges $E'' \subset E'$ according to the same missingness model f, along with all observed nonedges $V \times V - E'$ as negative examples (*SI Appendix*, section D). Unless other is specified, results reflect a choice of $\alpha = 0.8$ (i.e., 20% of edges are unobserved [holdout set]); other values produce qualitatively similar results.

Results

Prediction Error Diversity. If all link predictors exploit a common underlying signal of missingness, then one or a few predictors will consistently perform best across realistic inputs. Optimal link prediction could then be obtained by further leveraging this universal signal. In contrast, if different predictors exploit distinct signals, they will exhibit a diversity of errors in the form of heterogenous performance across inputs. In this case, there will be no best or worst method overall, and optimal link predictions can only be obtained by combining multiple methods. This dichotomy also holds at the level of predictor families, one of which could be best overall (e.g., topological methods), even if no one family member is best.

To distinguish these possibilities, we characterize the empirical distribution of errors by training a random forest classifier over the 203 link predictors applied to each of the 550 real-world networks (*SI Appendix*, section E). In this setting, the character of a predictor's errors is captured by its learned Gini importance (mean decrease in impurity) (11) within the random forest: the higher the Gini importance, the more generally useful the predictor is for correctly identifying missing links on that network. We then aggregate these importances across networks within a domain using a rank-weighted cross-entropy Monte Carlo (34) algorithm. If all methods exploit a common missingness signal (one method to rule them all), the same few predictors or predictor family will be assigned consistently greater importance across networks and domains. However, if there are multiple distinct signals (a diversity of errors), the learned importances will be highly heterogeneous across inputs, and no predictor or family will be best.

Across networks and domains, we find wide variation in individual and family-wise predictor importances, such that no individual method and no family of methods are best, or worst, on all networks. On individual networks, importances tend to be highly skewed, such that a relatively small subset of predictors accounts for the majority of prediction accuracy (*SI Appendix*, Fig. S2 and Table S4). However, the precise composition of this subset varies widely across both networks and families (*SI Appendix*, Figs. S3 and S4 and Tables S4 and S5), implying a broad diversity of errors and multiple distinct signals of missingness. At the same time, not all predictors perform well on realistic inputs (e.g., a subset of topological methods generally receives low importances), and most embedding-based predictors are typically mediocre. Nevertheless, each family contains some members that are ranked among the most important predictors for many, but not all, networks.

Across domains, predictor importances cluster in interesting ways, such that some individual and some families of predictors perform better on specific domains. For instance, examining the 10 most important predictors by domain (28 unique predictors; Fig. 1), we find that topological predictors, such as common neighbors or localized random walks, as well as distance-based embedding predictors, such as a Euclidean distance, perform well on social networks but less well on networks from other domains. In contrast, model-based methods perform relatively well across domains but often perform less well on social networks than do topological measures and some embedding-based methods. Together, these results indicate that predictor methods exhibit a broad diversity of errors, which tend correlate somewhat with scientific domain.

This performance heterogeneity implicates the practical relevance for link prediction of the No Free Lunch theorem (20), which proves that across all possible inputs, every machine learning method has the same average performance, and hence, accuracy must be assessed on a per input set basis. The observed diversity of errors indicates that none of the 203 individual predictors are a universally best method for the subset of all inputs that are realistic. However, that diversity also implies that a nearly optimal link prediction method for realistic inputs could be constructed by combining individual methods so that the best individual method is applied for each given input. Such a metalearning algorithm cannot circumvent the No Free Lunch theorem (*SI Appendix*, Figs. S10 and S11), but it can achieve optimal performance on realistic inputs by effectively redistributing its worse than average performance onto unrealistic inputs, which are unlikely to be encountered in practice. In the following sections, we develop and investigate the near-optimal performance of such an algorithm.

Stacking on Networks with Known Structure. Model "stacking" is a metalearning approach that learns to apply the best individual predictor according to the input's characteristics (25). Here, we assess the accuracy of model stacking both within and across families of prediction methods, which adds seven more prediction algorithms to our evaluation set.

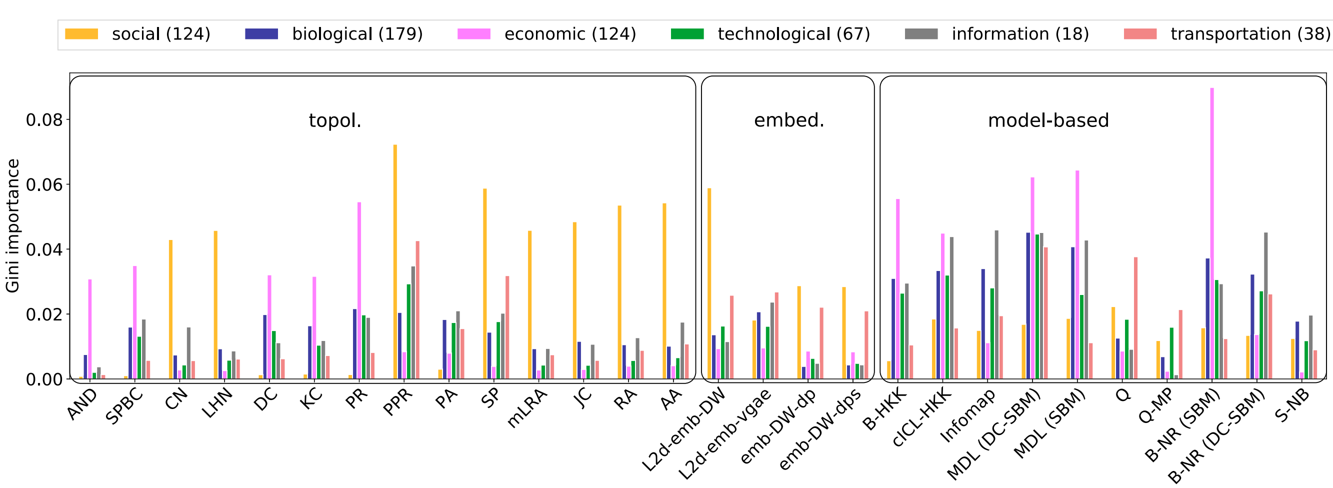

Fig. 1. The Gini importances for predicting missing links in networks within each of six scientific domains, for the 28 most important predictors, grouped by family, under a random forest classifier trained over all 203 predictors. Across domains, predictors exhibit widely different levels of importance, indicating a diversity of errors, such that no predictor is best overall. Here, topological predictors include shortest-path betweenness centrality (SPBC), common neighbors (CNs), Leicht–Holme–Newman index (LHN), personalized page rank (PPR), shortest path (SP), the mean neighbor entries within a low rank approximation (mLRA), Jaccard coefficient (JC), and the Adamic–Adar index (AA); embedding predictors include the L2 distance between embedded vectors under emb-DW (L2d-emb-DW) and the dot product (emb-vgae-dp) of embedded vectors under emb-vgae; and model-based predictors include Infomap (Infomap), stochastic block models with (MDL (DC-SBM), B-NR (DC-SBM)) and without degree corrections (MDL (SBM), B-NR (SBM)) and modularity (Q). (A complete list of abbreviations is given in *SI Appendix*, section A.)

Because the optimality of an algorithm's predictions can only be assessed when the underlying data-generating process is known, we first characterize the accuracy of model stacking using synthetic networks with known structure, for which we calculate an exact upper bound on link prediction accuracy (*SI Appendix*, section B). To provide a broad range of realistic variation in these tests, we use a structured random graph model, in which we systematically vary its degree distribution's variance, the

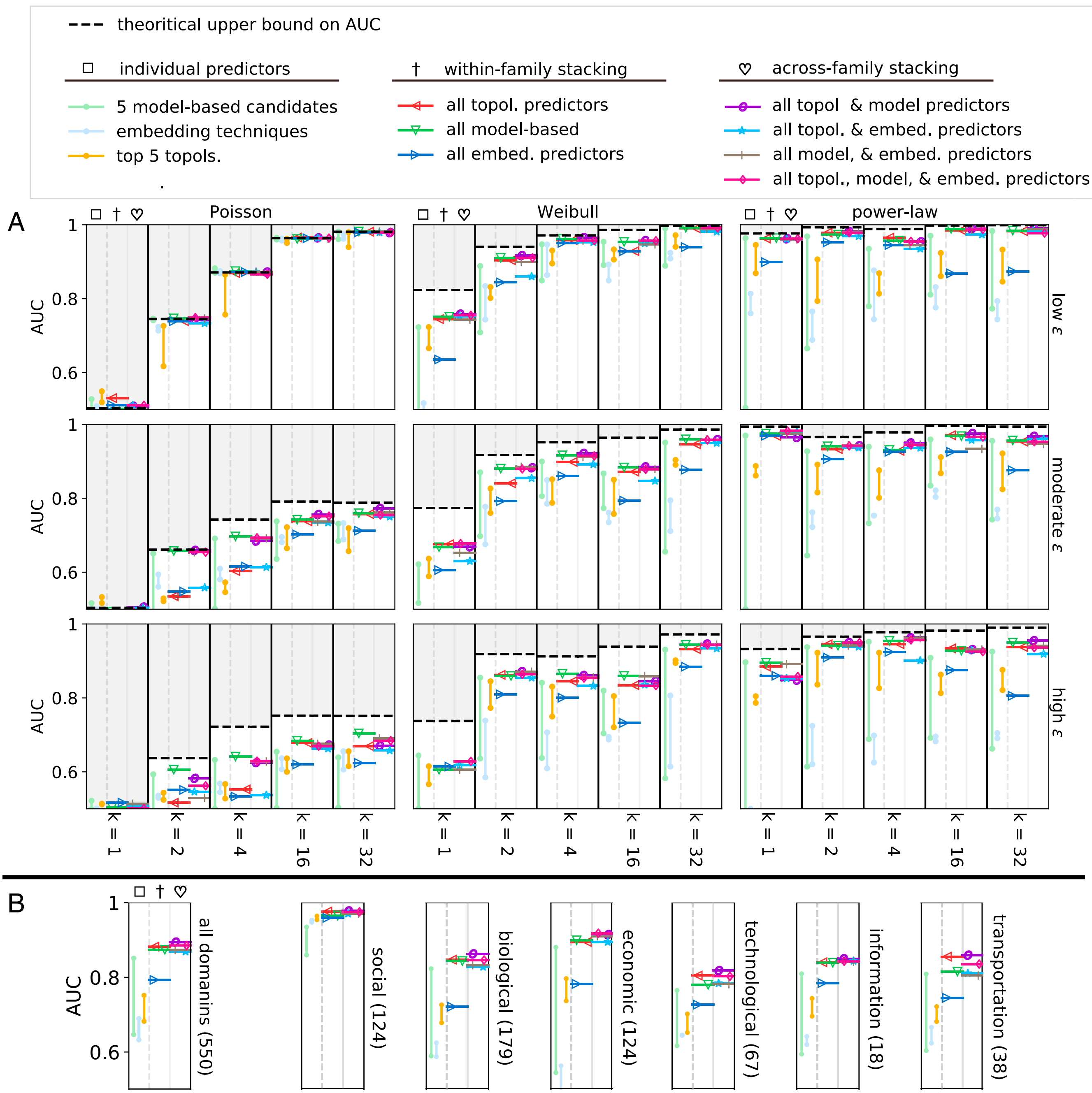

Fig. 2. (*A*) On synthetic networks, the mean link prediction performance (AUC) of selected individual predictors and all stacked algorithms across three forms of structural variability: degree distribution variability, from (*Left*) low (Poisson) to (*Right*) high (power law); fuzziness of community boundaries, ranging from (*Top*) low to (*Bottom*) high ($\epsilon = m_{out}/m_{in}$, the ratio of edges between the clusters to edges inside the clusters); and (from left to right within each subpanel) the number of communities k. Across settings, the dashed lines represent the theoretical maximum performance achievable by any link prediction algorithm (*SI Appendix*, section B). In each instance, stacked models perform optimally or nearly optimally and generally perform better when networks exhibit heavier-tailed degree distributions and more communities with distinct boundaries. *SI Appendix*, Table S11 lists the top five topological predictors for each synthetic network setting, which vary considerably. (*B*) On real-world networks, the mean link prediction performance for the same predictors over all domains and by individual domain. Both overall and within domains, stacked models exhibit superior performance, particularly the across-family versions, and they achieve nearly perfect accuracy on social networks. Performance varies considerably across individual domains, with biological and technological networks exhibiting the lowest link predictability. More complete results for individual topological and model-based predictors are given in *SI Appendix*, Figs. S8 and S9. For ease of interpretability, each panel's results are partitioned into three columns, showing (□) the performance range for selected individual predictors in each family (in the legend), (†) the results for within-family stacking, and (♡) the results for across-family stacking.

number of communities k, and the fuzziness of those community boundaries ϵ.

Across these structural variables, the upper limit on link predictability varies considerably (Fig. 2*A*), from no better than chance in a simple random graph ($k = 1$; Poisson) to nearly perfect in networks with many distinct communities and a power-law degree distribution. Predictability is lower (no methods can do well) with fewer communities (low k) or with more fuzzy boundaries (high ϵ) but higher with increasing variance in the degree distribution (Weibull or power law) or with dense clusters (low ϵ). Most methods, whether stacked or not, perform relatively well when predictability is low. However, as potential predictability increases, methods exhibit considerable dispersion in their accuracy, particularly among topological and embedding-based methods.

Regardless of the synthetic network's structure, we find that stacking methods are typically among the most accurate prediction algorithms, and they often achieve optimal or nearly optimal accuracy (Fig. 2*A*). For instance, the practical performance of the best stacked models is significantly closer to optimality than is the best on average individual predictor (all model-based or all topol., model, & embed., $\Delta\mathrm{AUC} = 0.04$ vs. MDL (DC-SBM), $\Delta\mathrm{AUC} = 0.07$; paired t test, $P < 10^{-4}$; *SI Appendix*, Tables S8 and S9), and they are far better than the average nonstacked topological and model-based methods ($\langle\Delta\mathrm{AUC}\rangle = 0.22$). We further note that here, the good performance of the MDL (DC-SBM) individual predictor is expected, as the synthetic networks are generated using a DC-SBM model (*SI Appendix*, section H and Table S10).

Stacking on Real-World Networks. To characterize the real-world accuracy of model stacking, we apply these methods and the individual predictors to our corpus of 550 real-world networks. We analyze the results within and across scientific domains and as a function of network size.

Across all networks and within individual domains, model stacking produces the most accurate predictions of missing links (Fig. 2*B* and Table 1), while some individual predictors also perform relatively well, particularly model-based ones. Applied to all networks, the average performances of the best stacked models are slightly but significantly better than the average performance of the best individual method (all topol. & model, $\langle\mathrm{AUC}\rangle = 0.89 \pm 0.09$, and all topol., model & embed., $\langle\mathrm{AUC}\rangle = 0.89 \pm 0.1$ vs. MDL (DC-SBM), $\langle\mathrm{AUC}\rangle = 0.85 \pm 0.11$; t test, $P < 10^{-12}$) and far better than the average performance of individual topological or model-based predictors ($\langle\mathrm{AUC}\rangle = 0.64$; Table 1 and *SI Appendix*, Table S6).

Stacking also achieves substantially better precision in its predictions (Table 1), which can be a desirable property in practice. In this particular experiment, the supervised learning step optimized the standard F measure over the holdout test set (*SI Appendix*, section A). Learning an optimal threshold via cross-validation produces nearly identical performance on our test corpus, and optimizing the AUC itself produces similar results but with slightly higher AUC scores (*SI Appendix*, Table S19).

Among the stacked models, the highest accuracy on real-world networks is achieved by stacking model-based and topological predictor families. Adding embedding-based predictors does not significantly improve accuracy, suggesting that the network embeddings do not capture more structural information than is represented by the model-based and topological families. This behavior aligns with our results on synthetic networks above, where the performances of stacking all predictors and stacking only model-based and topological predictors were nearly identical (*SI Appendix*, Table S8).

Applied to individual scientific domains, we find considerable variation in missing link predictability, which we take to be approximated by the most accurate stacked model (Fig. 2*B*). In particular, most predictors, both stacked and individual (*SI Appendix*, Figs. S8 and S9), perform well on social networks, and on these networks, model stacking achieves nearly perfect link prediction (up to $\mathrm{AUC} = 0.98 \pm 0.06$; *SI Appendix*, Table S13). In contrast, this upper limit is substantially lower in nonsocial domains, being lowest for technological networks ($\mathrm{AUC} = 0.82 \pm 0.09$; *SI Appendix*, Table S16), while marginally higher for biological, information, and transportation networks ($\mathrm{AUC} = 0.86$; *SI Appendix*, Tables S14, S17, and S18) and much higher for economic networks ($\mathrm{AUC} = 0.92 \pm 0.05$; *SI Appendix*, Table S15).

Stacked models also exhibit superior link prediction performance across real-world networks of different scales (number of edges m; Fig. 3) and tend to become more accurate as network size increases, where link prediction is inherently harder. We note, however, that on small networks ($m < 100$), a simple majority vote among model-based predictors slightly outperforms all stacking methods, while performing substantially worse than the best stacked model on larger networks ($m > 200$). Embedding-based methods perform poorly at most scales, but worst on smaller networks, suggesting a tendency to overfit (*SI Appendix*, Figs. S10 and S11). We note that stacking within that family does produce higher accuracies on larger networks but still lower than other stacked models.

These stacked models can also be extended to include other ensemble methods as individual predictors. For instance, Bayesian model averaging can be straightforwardly applied to several of the model-based predictors (29), including most variants of the stochastic block model (17). Applying this approach to the MDL (DC-SBM) individual predictor (9), we find that the averaged version tends to perform better, on average, than the unaveraged version (Fig. 3), although not on every network. However, we also find that including both the averaged

Table 1. Link prediction performance (mean ± SD) on holdout test set, measured by AUC, precision, and recall, for algorithms applied to the 550 structurally diverse networks in our corpus

Algorithm	AUC	Precision	Recall
Q	0.7 ± 0.15	0.06 ± 0.12	0.25 ± 0.28
Q-MR	0.67 ± 0.15	0.08 ± 0.13	0.31 ± 0.31
Q-MP	0.65 ± 0.14	0.04 ± 0.09	0.22 ± 0.25
B-NR (SBM)	0.81 ± 0.13	0.16 ± 0.25	0.23 ± 0.22
B-NR (DC-SBM)	0.71 ± 0.19	0.19 ± 0.23	0.21 ± 0.21
cICL-HKK	0.8 ± 0.13	0.24 ± 0.32	0.25 ± 0.25
B-HKK	0.78 ± 0.13	0.12 ± 0.22	0.21 ± 0.23
Infomap	0.74 ± 0.14	0.11 ± 0.14	0.29 ± 0.25
MDL (SBM)	0.81 ± 0.15	0.21 ± 0.29	0.24 ± 0.24
MDL (DC-SBM)	0.85 ± 0.11	0.13 ± 0.17	0.21 ± 0.18
S-NB	0.72 ± 0.18	0.32 ± 0.35	0.26 ± 0.24
Mean model-based	0.75 ± 0.16	0.15 ± 0.24	0.24 ± 0.25
Mean indiv. topol.	0.61 ± 0.14	0.09 ± 0.2	0.23 ± 0.27
Mean indiv. topol. & model	0.64 ± 0.15	0.1 ± 0.21	0.23 ± 0.27
emb-DW	0.63 ± 0.23	0.11 ± 0.17	0.3 ± 0.29
emb-vgae	0.69 ± 0.19	0.15 ± 0.21	0.25 ± 0.23
All topol.	0.88 ± 0.1	0.31 ± 0.33	0.35 ± 0.29
All model-based	0.87 ± 0.1	0.25 ± 0.28	0.29 ± 0.22
All embed.	0.79 ± 0.14	0.18 ± 0.23	0.27 ± 0.23
All topol. & model	0.89 ± 0.09	0.31 ± 0.34	0.34 ± 0.28
All topol. & embed.	0.87 ± 0.11	0.27 ± 0.31	0.33 ± 0.25
All model & embed.	0.87 ± 0.11	0.23 ± 0.26	0.3 ± 0.23
All topol., model, & embed.	0.89 ± 0.1	0.28 ± 0.31	0.34 ± 0.26

A complete list of abbreviations is given in *SI Appendix*, section A.

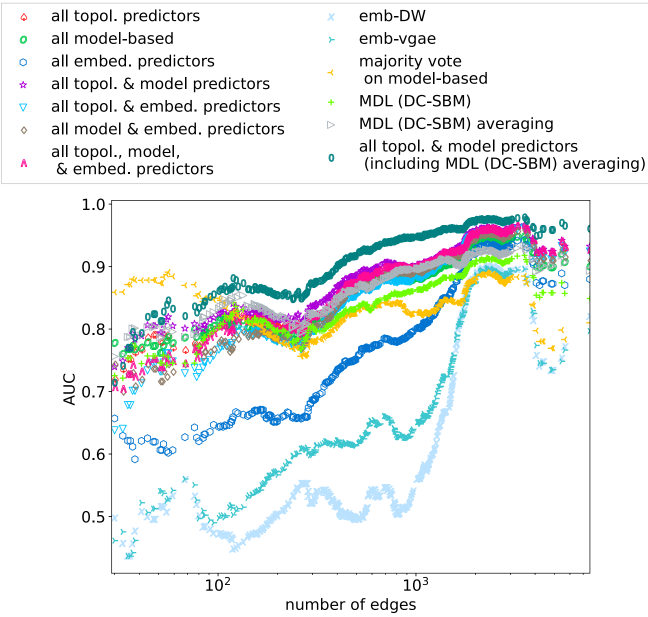

Fig. 3. Mean link prediction performance (AUC) as a function of network size (number of edges *m*) for stacked models and select individual predictors, applied to 550 real-world networks, smoothed using a sliding window. Overall, stacking topological predictors, model-based predictors, or both yield superior performance but especially on larger networks where link prediction is inherently more difficult. (A complete list of abbreviations is given in *SI Appendix*, section A.)

and unaveraged versions in the topological and model-based stacked model further improves its overall link prediction accuracy (Fig. 3; *SI Appendix*, section H, Fig. S18, and Table S24 have more details). This behavior indicates that the averaged version itself makes some errors that are distinct from those of the unaveraged version, and stacking is able to learn to exploit both.

Sufficiency and Optimality. In practice, the optimality of a metalearning method can only be established indirectly, over a set of considered predictors applied to a sufficiently diverse range of empirical tests cases (20). We assess this indirect evidence for stacked link prediction models through two numerical experiments.

In the first, we consider how performance varies as a function of the number of predictors stacked, either within or across families. Evidence for optimality here appears as an early saturation, in which performance achieves its maximum prior to the inclusion of all available individual predictors. This behavior would indicate that a subset of predictors is sufficient to capture the same information as the total set. To test for this early-saturation signature, we first train a random forest classifier on all predictors in each of our stacked models and calculate each predictor's within-model Gini importance. For each stacked model, we then build a new sequence of submodels in which we stack only the k most important predictors at a time and assess its performance on the test corpus.

In each of the stacked models, performance exhibits a classic saturation pattern: it increases quickly as the 10 most important predictors are included and then stabilizes by around 30 predictors (Fig. 4 and *SI Appendix*, Fig. S5). Performance then degrades slightly beyond 30 to 50 included predictors, suggesting a slight degree of overfitting in the full models. Notably, each within- and across-family model exhibits a similar saturation curve, except for the embedding-only model, which saturates early and at a lower level than other stacked models. This similar behavior suggests that these families of predictors are capturing similar missingness signals, despite their different underlying representations of the network structure. As in other experiments, the best saturation behavior is achieved by stacking model-based and topological predictors.

In the second, we evaluate whether individual predictors represent "weak" learners in the sense that their link prediction performance is better than random. In general, we find that nearly all of the predictors satisfy this condition (*SI Appendix*, Figs. S6 and S7), implying that they can be combined according to the Adaboost theorem to construct an optimal algorithm (35). Replacing the random forest algorithm within our stacking approach with a standard boosting algorithm also produces nearly identical performance on our test corpus (*SI Appendix*, Tables S20–S23). The similar performance between the two methods suggests that relatively little additional performance is likely possible using other metalearning approaches over the same set of predictors.

Discussion

Developing more accurate methods for predicting missing links in networks would help reduce the use of scarce resources in collecting network data and would provide more powerful tools for evaluating and comparing network models of complex systems. The literature on such methods gives an unmistakable impression that most published algorithms produce reasonably accurate predictions. However, relatively few of these studies present systematic comparisons across different families of methods, and they typically draw their test cases from a narrow set of empirical networks (e.g., social networks). As a result, it has remained unknown whether a single best predictor or family of predictors exists, how link predictability itself varies across different methods and scientific domains, or how close to optimality current methods may be.

Our broad analysis of individual link prediction algorithms, representing three large and popular families of such methods, applied to a large corpus of structurally diverse networks shows definitively that common predictors in fact exhibit a broad diversity of errors across realistic inputs (Fig. 1 and *SI Appendix*, Fig. S2). Moreover, this diversity is such that no one predictor, and no family of predictors, is overall best, or worst, in practice

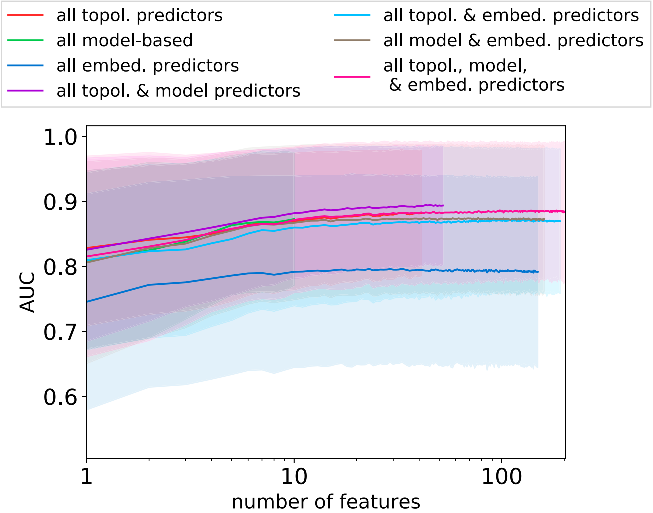

Fig. 4. Mean link prediction performance (AUC) as a function of the number of stacked features, for within- and across-family stacked models, applied to 550 real-world networks. The shaded regions show the SD, and the early-saturation behavior (at between 10 and 50 predictors) indicates that a small subset of predictors is sufficient to capture the same information as the total set.

(*SI Appendix*, Figs. S3, S10, and S11 and Table S4). The common practice of evaluating link prediction algorithms using a relatively narrow range of test cases is thus problematic. The far broader range of empirical networks and algorithms considered here shows that, generally speaking, good performance on a few test cases does not generalize across inputs. The empirical diversity of errors we find implicates the practical relevance of the No Free Lunch theorem (20) for predicting missing links in complex networks and suggests that optimal performance on realistic inputs may only be achieved by combining methods (e.g., via metalearning) to construct an ensemble whose domain of best performance matches the particular structural diversity of real-world networks.

Model stacking is a popular metalearning approach, and our results indicate that, after adapted to a network setting, it can produce highly accurate predictions of missing links by combining either topological predictors alone, model-based predictors alone, or both. Applied to structurally diverse synthetic networks, for which we may calculate optimal performance, stacking achieves optimal or near-optimal accuracy, and accuracy is generally closer to perfect when networks exhibit a highly variable degree distribution and/or many, structurally distinct communities (Fig. 2*A*).

Similarly, applied to empirical networks, stacking produces more accurate predictions across inputs than any individual predictor (Fig. 2*B* and Table 1), and these predictions appear to be nearly optimal (i.e., we find little evidence that further accuracy can be achieved using this set of predictors; Fig. 4), even with alternative metalearning approaches. Of course, it remains possible that accuracy could be further improved by incorporating specific new predictors or new families within the stacked models, if they provide better prediction coverage of some input networks than the present set of predictors. For instance, in an experiment suggested by a reviewer, we find that incorporating a predictor based on Bayesian model averaging the MDL (DC-SBM) improves the stacked model's overall link prediction performance on empirical networks (Fig. 3 and *SI Appendix*).

Beyond its strong predictive utility for missing links, model stacking also provides a useful framework for understanding why predictors make different errors on different inputs. Although a full investigation and interpretation of these differences are beyond the scope of the investigation here, our findings do provide suggestive preliminary results. For instance, we find that 1) stacking a small number of topological predictors, including personalized page rank, degree product, and shortest path count, can yield equivalent performance to a more sophisticated and flexible predictor like MDL (DC-SBM) (*SI Appendix*, Fig. S13); 2) topological predictors that model social processes like triadic closure and homophily (e.g., local clustering, Jaccard coefficient, and degree assortativity) are particularly effective in social networks, the context for which they were designed, but perform worse in other contexts (e.g., biological networks) (*SI Appendix*, section G); 3) linear regression can be applied over specific topological features to generate more interpretable hypotheses about the structure of networks (*SI Appendix*, Fig. S17); and 4) performance can depend on network size, such that individual predictors can outperform stacked models on smaller networks, where data for learning a stacked model are in short supply (*SI Appendix*, Figs. S14–S16).

Across networks drawn from different scientific domains (e.g., social vs. biological networks), we find substantial variation in link predictor performance, both for individual predictors and for stacked models. This heterogeneity suggests that the basic task of link prediction may be fundamentally harder in some domains of networks than others. Most algorithms produce highly accurate predictions in social networks, which are stereotypically rich in triangles (local clustering), exhibit broad degree distributions, and are composed of assortative communities, suggesting that link prediction in social networks may simply be easier (36) than in nonsocial network settings. In fact, stacked models achieve nearly perfect accuracy at distinguishing true positives (missing links) from true negatives (nonedges) in social networks (Fig. 2*B* and *SI Appendix*, Table S13). An alternative interpretation of this difference is that the existing families of predictors exhibit some degree of selective inference (i.e., they work well on social networks because social network data are the most common inspiration and application for link prediction methods). Our results make it clear that developing more accurate individual predictors for nonsocial networks (e.g., biological and informational networks) is an important direction of future work. Progress along these lines will help clarify whether link prediction is fundamentally harder in nonsocial domains and why.

Across our analyses, embedding-based methods, which are instances of representation learning on networks, generally perform more poorly than do either topological or model-based predictors. This behavior is similar to recent results in statistical forecasting, which found that neural network and other machine learning methods perform less well by themselves than when combined with other, conventional statistical methods (37, 38). A useful direction of future work on link prediction would specifically investigate tuning embedding-based methods to perform better on the task of link prediction.

Recent theoretical work on model stacking for nonrelational data provides a strong justification for using stacking to combine models when the goal is out-of-sample prediction (27, 28). Moreover, these results suggest that in the practical setting in which the true data-generating model is unknown, model stacking learns an effective Bayes-optimal model of the input distribution. Extending these results to relational data would shed further light on the optimality of model stacking for link prediction and help us assess how close to optimality is their observed performance here.

Nevertheless, our findings suggest that stacking achieves nearly optimal performance across a wide variety of realistic inputs. It is likely that efforts to develop new individual link prediction algorithms will continue, and these efforts will be especially beneficial in specific application domains (e.g., predicting missing links in genetic regulatory networks or in food webs). Evaluations of new predictors, however, should be carried out in the context of metalearning, in order to assess whether they improve the overall prediction coverage embodied by the state-of-the-art stacked models applied to realistic inputs. Similarly, these evaluations should be conducted on a large and structurally diverse corpus of empirical networks, like the one considered here. More narrow evaluations are unlikely to produce reliable estimates of predictor generalization. Fortunately, stacked models can easily be extended to incorporate any new predictors as they are developed, providing an incremental path toward fully optimal predictions.

Data Availability. Network data and code for replication and reuse have been deposited in GitHub (https://github.com/Aghasemian/OptimalLinkPrediction).

ACKNOWLEDGMENTS. We thank David Wolpert, Brendan Tracey, Cristopher Moore, and Tiago Peixoto for helpful conversations, acknowledge the BioFrontiers Computing Core at the University of Colorado Boulder for providing High Performance Computing resources (NIH Grant 1S10OD012300) supported by BioFrontiers IT, and thank the Information Sciences Institute at the University of Southern California for hosting A. Ghasemian during this project. Financial support for this research was provided in part by NSF Grant IIS-1452718 (to A. Ghasemian and A.C.), Army Research Office Award W911NF-15-1-0259 (to A. Galstyan), NSF Grant IIS-1409177 (to E.M.A.), ONR Award YIP N00014-14-1-0485 (to E.M.A.), and ONR Award YIP N00014-17-1-2131 (to E.M.A.).

1. G. Kossinets, Effects of missing data in social networks. *Soc. Network.* **28**, 247–268 (2006).
2. M. Fire *et al.*, Computationally efficient link prediction in a variety of social networks. *ACM Trans. Intell. Syst. Technol. (TIST)* **5**, 10 (2013).
3. L. Lü, T. Zhou, Link prediction in complex networks: A survey. *Phys. Stat. Mech. Appl. A* **390**, 1150–1170 (2011).
4. M. Nagarajan *et al.*, "Predicting future scientific discoveries based on a networked analysis of the past literature" in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, 2015), pp. 2019–2028.
5. G. C. Kane, M. Alavi, G. J. Labianca, S. Borgatti, What's different about social media networks? A framework and research agenda. *MIS Q.* **38**, 274–304 (2014).
6. M. Burgess, E. Adar, M. Cafarella, Link-prediction enhanced consensus clustering for complex networks. *PLoS One* **11**, e0153384 (2016).
7. A. Mirshahvalad, J. Lindholm, M. Derlen, M. Rosvall, Significant communities in large sparse networks. *PLoS One* **7**, e33721 (2012).
8. A. Ghasemian, H. Hosseinmardi, A. Clauset, Evaluating overfit and underfit in models of network community structure. *IEEE Trans. Knowl. Data Eng.* **32**, 1722–1735 (2019).
9. T. Vallès-Català, T. P. Peixoto, M. Sales-Pardo, R. Guimerà, Consistencies and inconsistencies between model selection and link prediction in networks. *Phys. Rev. E* **97**, 062316 (2018).
10. S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection. *Stat. Surv.* **4**, 40–79 (2010).
11. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, NY, 2009).
12. A. Clauset, C. Moore, M. E. J. Newman, Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101 (2008).
13. V. Martínez, F. Berzal, J. C. Cubero, A survey of link prediction in complex networks. *ACM Comput. Surv.* **49**, 69 (2017).
14. M. Al Hasan, M. J. Zaki, "A survey of link prediction in social networks" in *Social Network Data Analytics*, C. C. Aggarwal, Ed. (Springer, New York, NY, 2011), pp. 243–275.
15. D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks. *J. Assoc. Inform. Sci. Technol.* **58**, 1019–1031 (2007).
16. T. Zhou, L. Lü, Y. C. Zhang, Predicting missing links via local information. *European Phys. J. B* **71**, 623–630 (2009).
17. R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 22073–22078 (2009).
18. A. Grover, J. Leskovec, "node2vec: Scalable feature learning for networks" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, 2016), pp. 855–864.
19. H. Cai, V. W. Zheng, K. C. C. Chang, A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.* **30**, 1616–1637 (2018).
20. D. H. Wolpert, W. G. Macready, No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82 (1997).
21. L. Peel, D. B. Larremore, A. Clauset, The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**, e1602548 (2017).
22. R. E. Schapire, The strength of weak learnability. *Mach. Learn.* **5**, 197–227 (1990).
23. L. Breiman, Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
24. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
25. D. H. Wolpert, Stacked generalization. *Neural Network.* **5**, 241–259 (1992).
26. R. E. Schapire, "A brief introduction to boosting" in *Proceedings of the 16th International Joint Conference on Artificial intelligence*, T. Dean, Ed. (Morgan Kaufmann Publishers Inc., San Francisco, CA, 1999), vol. 2, pp. 1401–1406.
27. T. Le, B. Clarke, A bayes interpretation of stacking for $\mathcal{M}$-complete and $\mathcal{M}$-open settings. *Bayesian Anal.* **12**, 807–829 (2017).
28. Y. Yao, A. Vehtari, D. Simpson, A. Gelman, Using stacking to average bayesian predictive distributions. *Bayesian Anal.* **13**, 917–1007 (2018).
29. J. A. Hoeting, D. Madigan, A. E. Raftery, C. T. Volinsky, Bayesian model averaging: A tutorial. *Stat. Sci.* **14**, 382–401 (1999).
30. P. Domingos, "Why does bagging work? A Bayesian account and its implications" in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, D. Heckerman, R. Uthurusamy, H. Mannila, D. Pregibon, Eds. (AAAI Press, 1997), pp. 155–158.
31. T. P. Minka, Bayesian model averaging is not model combination MIT Media Lab note (2000). https://tminka.github.io/papers/bma.html. Accessed 17 August 2020.
32. H. C. Kim, Z. Ghahramani, "Bayesian classifier combination" in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS*, N. D. Lawrence, M. Girolami, Eds. (Proceedings of Machine Learning Research, 2012), pp. 619–627.
33. Y. Koren, The BellKor solution to the Netflix Grand Prize. Netflix Prize Documentation 81, 1–10 (2009). https://netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf. Accessed 17 August 2020.
34. V. Pihur, S. Datta, S. Datta, Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach. *Bioinformatics* **23**, 1607–1615 (2007).
35. Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
36. A. Epasto, B. Perozzi, "Is a single embedding enough? Learning node representations that capture multiple social contexts" in *The World Wide Web Conference*, L. Liu, R. White, Eds. (Association for Computing Machinery, New York, NY, 2019), pp. 394–404.
37. S. Makridakis, E. Spiliotis, V. Assimakopoulos, The M4 competition: Results, findings, conclusion and way forward. *Int. J. Forecast.* **34**, 802–808 (2018).
38. S. Makridakis, E. Spiliotis, V. Assimakopoulos, Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS One* **13**, e0194889 (2018).